**Assignment Submission Cover Sheet**

|  |  |
| --- | --- |
| **Programme Title:** | **MSc Business Analytics** |
| **Module Code and Title:** | **BU7143 Business Data Mining** |
| **Assessment Title:** | **Group Report** |
| **Group Number:** | **Team 5** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Student Name and Contribution** | **%** |  | **%** |
| **1. Lim Yue Ying Veronica** |  | **4. Kai Kei Cheung** |  |
| **2. Temitayo Coker** |  | **5. Ashley Chew Bo Qing** |  |
| **3. Huang Xueyan** |  | **6. Yanliang Li** |  |

For group work – individual % contributions need to be stated **only** where they **are not equal.**

Please read the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at:  <http://www.tcd.ie/calendar>

Online Tutorial on avoiding plagiarism ‘Ready, Steady, Write’, is located at <http://tcd-ie.libguides.com/plagiarism/ready-steady-write>

CMS

Data Mining Report

**Executive summary**

As a trusted partner and steward, Centers for Medicare & Medicaid Services (CMS) is a federal organisation dedicated to advancing health equity, expanding coverage, and improving public health outcomes. The majority of expenses related to diabetes intervention, treatment, and prevention in the U.S. are covered by CMS (*CMS, 2021, “Blog” section*).

Healthcare, in general, has always been a large expenditure for the United States. In 2020, U.S. healthcare spending grew by 9.7% and reached an annual high of $4.1 trillion or $12,530 per person. Healthcare spending accounted for 19.7% of the U.S. Gross Domestic Product (GDP) and growth in national health expenditure is expected to reach $6.8 trillion by 2030, thus highlighting the immense scale of this problem (*CMS, 2020, “Historical” section*).

More specifically, diabetes is considered to be the most expensive chronic condition in the United States, and as such, has become a huge economic burden on both taxpayers and federal reserves (*ADA, 2017*). Type 2 diabetes, however, is a form of diabetes that can be avoided in the early stage with appropriate evidence-based lifestyle change programs or taking pre-treatment medication such as Metformin.

Furthermore, appropriate classification of patients into these classes can lead to higher conversion rates of patients into these assisted programs offered by the Centers for Disease Control and Prevention (CDC) such as the National Diabetes Prevention Program (NDPP) for preventing “*low-risk*” Type 2 pre-diabetic diagnosis and Diabetes Self-Management Education and Support (DSMES) services for preventing the complications arising from a “*high-risk*” Type 2 pre-diabetic diagnosis.

The appropriate classification of patients’ Type 2 pre-diabetic status into these 2 classes of “*low-risk*” and “*high-risk*” for placement into lifestyle change programs is the goal of this report. Early diagnosis can lead to lifestyle changes and more effective treatment, and this can be achieved through the Data Mining strategy that focuses on making predictive models for diabetes risks which will become an important tool for the public and public health officials (*Galaviz, 2015*). If this problem is properly handled, it not only aids in the early prevention of Type 2 diabetes, which may enhance the patient's general quality of life, but it also aims to minimise the tax burden on individual taxpayers and pharmacy costs covered by CMS.

The balanced dataset from the Behavioural Risk Factor Surveillance System (BRFSS) survey will be the core focus of this project with the key challenge being applying and implementing the appropriate model and seeing as the chosen dataset has an equal 50% split, this means that the dataset is not reflective of the entire population which could lead to the model over-predicting and producing more Type I errors. Further recommendations and a cost analysis will be presented as part of the findings of the project at the end of the report.

**Problem Description**

***Business Problem***

Diabetes is becoming more and more prevalent with diabetes-related medical expenses averaging $16,752 annually, of which $9,601 can be attributed to the condition, which is 2.3 times higher than those without the condition. While there are different types of diabetes, Type 2 diabetes is the most common form, and its prevalence varies by age, education, income, location, race, and other social determinants of health. As a result, much of the financial burden falls on those of lower socio-economic status which causes significant problems for patients themselves as well as CMS. CMS as a federal service, provides medical insurance coverage for inpatient hospital stays, doctors' services, outpatient care, medical supplies, preventative treatments, and so on - this places a massive burden on the economy, with diagnosed diabetes costing roughly $327 billion and total costs with undiagnosed diabetes and pre-diabetes approaching $400 billion annually (*National Diabetes Statistics Report, 2022*).

This trend is likely to continue with the CDC estimating that 1-in-5 diabetics, and roughly 8-in-10 pre-diabetics are unaware of their risk, as a result, this will put additional strain on already strained resources by the time patients’ become aware of their diagnoses and their options for treatment shift from lifestyle change programs to complete reliance on expensive medication and procedures that have to be covered by CMS.

It is important to recognise the scale of the diabetes-related healthcare spending to assess the relevance of the business problem - that is, the disproportionate usage of healthcare funding on Type 2 diabetes. In doing so, CMS can address the problem and through effective prediction of patients’ risk status, minimise patients’ reliance on CMS-covered drugs and treatments. Therefore, CMS can efficiently allocate healthcare funding and resources to treat other serious chronic illnesses plaguing the U.S. such as heart disease, arthritis, obesity, and stroke.

The key elements regarding the CMS business problem are listed below:

* **Strategy:** Effective prediction and classification of patients’ Type 2 diabetes status
* **Goal:** Reduce pharmacy costs and increase patient conversion into lifestyle change programs and services
* **Stakeholders:**
  + State and federal government agencies
  + Persons and families
  + Hospitals, pharmacies and clinicians
  + Health plan providers and academic researchers
* **Opportunities**:
  + Minimise tax burdens
  + Prevent serious Type 2 diabetes complications
  + Reduces patients’ risk of Type 2 diabetes
  + Improvements in quality of life

***Analytical Goal***

To effectively classify current and potential patients into 2 distinct classes, an array of data mining techniques will be applied. The analytical goal aims to increase patients’ awareness of their diabetes risk status so as to lower the likelihood of a diabetes diagnosis and delay the onset of significant complications. Thus, maximising the future value of CMS funding available to treat other chronic diseases and optimising patients’ present & future quality of life.

The 2 distinct classes are listed below:

1.   “***Low-risk***” pre-diabetic

2.   “***High-risk***” pre-diabetic

Hence, through the data mining solution, patients will be classified into one of these 2 classes and offered an affordable, evidence-based approach for lifestyle change that prevents diabetes and its related complications.

**Data Description**

The CDC conducts an annual telephone survey on health-related topics called the Behavioural Risk Factor Surveillance System (BRFSS) with over 400,000 Americans participating each year providing information on health-related risk behaviours, chronic health issues, and the use of preventative treatments. The original dataset contains 330 features based on questions directly asked from respondents, or calculated variables based on individual participant responses; and 441,455 responses from respondents. The dataset contains 3 sets of cleaned data each with 21 features variables described in ***Appendix A***, 1 target variable with 2 - 3 classes, and varying degrees of class balance or imbalance.

Within the BRFSS dataset, a cleaned file of 70,692 survey responses with an equal 50-50 split of respondents with no diabetes and with either pre-diabetes or diabetes will be the focus of the data mining techniques. The target variable, ***Diabetes\_binary***, has 2 classes: 0 is for no diabetes, and 1 is for pre-diabetes or diabetes. This dataset has 21 feature variables of both binary and integer data types and is balanced as seen in ***Figure 1*** below.

表格

描述已自动生成

表格

描述已自动生成

**Figure 1:** Head of Behavioural Risk Factor Surveillance System (BRFSS) dataset

**Data Preparation and Visualisation**

Pre-processing steps on the dataset included data partitioning into a training set (60%) and a validation set (40%) to improve the scalability of the models and avoid overly optimistic estimates of the model’s accuracy. As the dataset is already cleaned with no null or missing values, the process of data cleaning and variable transformation was not necessary. Subsequently, the partitioned datasets were normalised by rescaling the data using the *preProcess* function from the caret package in R to prepare the dataset(s) for analyses.

**Data Mining Solution**

​​The comparative evaluation for this project includes a trade-off between accuracy, sensitivity, and specificity. To build an effective predictive model to classify patients' Type 2 diabetes’ status as either “***low-risk”*** or “***high-risk”*** pre-diabetic, a collection of both descriptive and predictive models were created and analysed below:

* Principal Component Analysis (PCA)
* k-means clustering
* Linear regression model
* Logistic regression
* k-Nearest Neighbours (KNN)
* Naïve Bayes
* Classification trees
* Random Forest
* Boosted Tree
* Majority Ensemble
* Average Ensemble

The Diabetes Health Indicators Dataset has a large number of variables which could lead to the overfitting of the chosen model to the data as well as violating the assumptions of whichever modelling tactic is being used. As a result, in order to understand the relationship(s) between variables, it is necessary to reduce the dimensions of the feature space. This dimensionality reduction can be achieved through the means of *feature extraction*, where the 21 feature variables are combined in a specific way into *components* and ordered according to their “predictive power” with regards to the target variable, whilst dropping the “least important” variables and still retaining the most valuable parts of all the variables.

According to the scree plot in ***Figure 2*** below, the optimal number of components to include in the model is 3, as they are capturing a majority (~ 32%) of the variance in the dataset.

图片包含 图形用户界面

描述已自动生成

**Figure 2:** Scree plot of principal components

However, due to the balanced nature of the dataset, all variables are considered important towards the overall performance and accuracy of the chosen model(s), as removal of any variable could unfavourably skew the outcome of the dataset towards a patient’s diabetes status being “*low-risk*” or “*high-risk*”.

Seeing as the Principal Component Analysis (PCA) cannot be used to drop variables in this particular dataset, it was used in conjunction with the classification trees to better understand the relationships between the variables in each of the 3 components that explains the majority of the variance in the dataset. The loadings plots (***Figures 3 – 5***) and variable importance scores from the pruned classification tree (***Table 1***) shows that the 3 principal components (PC1, PC2, and PC3) can be further classified as 3 clusters named, and categorised as follows:

**PC1 / Cluster 1**

Patient’s ***Health Status***: GenHlth, HighBP, BMI, PhysHlth, HighChol, DiffWalk, HeartDiseaseorAttack, MentHlth, PhysActivity, Stroke, and Smoker

This first cluster or principal component contains variables that describes a patient’s overall health condition including their general, physical, and mental health; blood pressure and cholesterol levels; history of coronary heart disease (CHD), myocardial infarction (MI), and stroke; difficulty walking or climbing stairs; and body mass index measures.

**PC2 / Cluster 2**

Patient’s ***Lifestyle***: Age, Income, Education, Sex, CholCheck, NoDocbcCost, and AnyHealthcare

This second cluster or principal component contains variables that describes a patient’s lifestyle including their age, income, education, sex, as well as information surrounding their financial access to healthcare services and facilities.

**PC3 / Cluster 3**

Patient’s ***Consumption Habits***: HvyAlcoholConsump, Fruits, and Veggies

This third cluster or principal component contains variables that describe a patient’s consumption habits to include what they eat and drink.

表格

描述已自动生成

**Table 1:** Variable importance of variables in pruned classification tree

图表

描述已自动生成

**Figure 3:** Loadings Plot of PC1/Cluster 1 – *Patient’s Health Status*

*图表

描述已自动生成*

**Figure 4:** Loadings Plot of PC2/Cluster 2 – *Patient’s Lifestyle*

图表, 瀑布图

描述已自动生成

**Figure 5:** Loadings Plot of PC3/Cluster 3 – *Patient’s Consumption Habits*

Through this new classification of principal components as clusters, CMS can better understand the factors that put patients at risk of becoming pre-diabetic and exacerbate the complications for patients that are already diabetic. As a result, CMS is able to augment the existing evidence-based lifestyle change programs to reflect these factors so as to delay the onset of Type-2 diabetes and the self-management of its related complications more effectively.

In terms of accuracy and sensitivity, the pruned classification tree performed the best, according to the analysis of the models, with a strong inclination toward Type I errors (false positives), as shown in ***Table 2*** below. As a result, CMS and its stakeholders can use this predictive model to predict Type 2 diabetes-related health events and screen those who are at "*high-risk*." Although causal modelling may and is commonly used to identify risk factors, predictive modelling employing classification trees offers highly relevant information for individualised Type 2 diabetes risk projections as well as for guiding treatment regimens. Physicians, counsellors, health educators, politicians, and other stakeholders can frequently find significant value in this predictive knowledge, which they can use to recommend course correction or interventions like NDPP and DSMES services.

**表格

描述已自动生成**

**Table 2:** Summary of model performance metrics

**Implementation**

Implementing the model requires an understanding of the costs associated with each decision made and here, there are two considerations:

1. Over-predict patients with diabetes
2. Misclassify the patients as low-risk diabetics

If CMS chooses to overpredict, the costs associated with the decision would be the cost of the program per patient. While if CMS chooses to misclassify patients as low-risk diabetics, the cost would be the cost of the pharmaceutical drugs. Choosing to misclassify patients could potentially pose a threat to patients’ health as well as higher costs associated with the medical expenses for treatment. If diabetes is not treated and diagnosed at an early stage, patients could face an increased risk of diabetes complications such as nerve, eye, and kidney damage, as well as diabetic ketoacidosis. As mentioned earlier, medical expenses for individuals with diabetes are 2.3 times higher than those without, thus it is better for CMS to over-estimate than under-estimate patients with diabetes to reduce overall costs for medical expenses. To put this in perspective, it will cost CMS $500 per year (~ $41.70 per month) (*NDPP, n.d*.) to enrol a patient in a lifestyle change program as opposed to paying $232 per month (*Lindberg, 2020*) on diabetes drugs - generating cost savings of 82% annually.

The proposed decision to over-predict is also supported by an understanding of the cost of total healthcare expenses for patients’ lifetimes. The lifetime expenses for early diagnoses will range between $70,560 - $96,333 and for late diagnosis, between $120,415 - $144,856 (*Ou et al., 2016*). Although the expenditures associated with diabetes were the most significant factor, an early diagnosis often results in lower lifetime health care costs as well as improves patients’ overall health.

**Recommendations**

The classification model built can be used by CMS to effectively predict and identify the risk status of diabetes patients. In addition, this model can be used as a basis for further development to provide a better classification model that can accurately identify the varying severity of diabetes within patients. In doing so, CMS would be able to place the identified patients into relevant programs to prevent further progression of diabetes complications.

Based on the classification, patients' can be recommended and placed into the following existing lifestyle change programs offered by the CDC in collaboration with CMS:

1. “ *Low-risk*” pre-diabetes - National Diabetes Prevention Program (National DPP)
2. “*High-risk*” pre-diabetes - Diabetes Self-Management Education and Support (DSMES) services

With the programs in place to help patients build healthy lifestyle habits and improve their overall health, this reduces the number of patients with diabetes in the long-term. Doing so would reduce the funding that CMS would potentially use for pharmacy claims by patients with diabetes, which achieves the goal of the business strategy outlined earlier in the report. The reduction of diabetes-related pharmacy claims frees up funding for CMS to treat patients with other chronic diseases - thereby addressing the business problem set out in the report.

**Conclusion**

This report aims to address CMS’ business problem to accurately identify “*low-risk*” and “*high-risk*” diabetic patients. Therefore, the chosen classification model was evaluated with a focus on the variable's predictive power and sensitivity of the model.

Of the predictor models developed for this report, while some have poor performance or similar performance in their accuracy and sensitivity, the classification tree model had the best performance in its accuracy and sensitivity. Therefore, it was selected as the predictor model for CMS to implement. However, it is important to note that this model developed, and the data used does not come without its limitations.

**Advantages and Limitations**

The classification tree model is costly and time-consuming to implement which might be a strong point of contention for CMS, especially as a government agency trying to efficiently allocate and reduce disproportionate usage of funding. However, considering the long-term benefits of preventing Type 2 diabetes and its complications, this model is a good investment for CMS.

Another consideration is that the dataset used to develop and train the classification tree model is biased with an equal 50% split of respondents with no diabetes and with either pre-diabetes or diabetes. Therefore, it is not a true representation of the population, which skews the distribution of the prediction to over-predict the patients with diabetes. Despite such limitations, the prediction model developed is still a useful tool and model for CMS to address its business problem.

**Appendix**

**Appendix A**

表格

描述已自动生成

**References**

American Diabetes Association. (2017) *The Cost of Diabetes*. Diabetes Care. <https://diabetes.org/about-us/statistics/cost-diabetes>

Brooks-LaSure, B. (2021, September 9), *My First 100 Days and Where We Go From Here: A Strategic Vision for CMS*. Centers for Medicare & Medicaid Services. Retrieved December 9, 2022, from <https://www.cms.gov/blog/my-first-100-days-and-where-we-go-here-strategic-vision-cms>

Centers for Disease Control and Prevention (2022, June 29). *National Diabetes Statistics Report*. Retrieved December 9, 2022, from [https://www.cdc.gov/diabetes/data/statistics-report/index.html](https://www/cdc.gov/diabetes/data/statistics-report/index.html)

Centers for Disease Control and Prevention (2022, October 25). *By the Numbers: Diabetes in America*. Retrieved December 9, 2022, from <https://www.cdc.gov/diabetes/health-equity/diabetes-by-the-numbers.html#:~:text=Cost%20of%20Diabetes,%249%2C601%20per%20person%20person%20in%202017>.

Centers for Medicare & Medicaid Services. *National Health Expenditure Data - Historical*. Retrieved December 9, 2022, from <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical>

Centers for Disease Control and Prevention (2022, September 30). *Diabetes Self-Management Education and Support (DSMES) Toolkit*. Retrieved December 9, 2022, from <https://www.cdc.gov/diabetes/dsmes-toolkit/index.html>

Centers for Disease Control and Prevention (2022, October 30). *National Diabetes Prevention Program*. Retrieved December 9, 2022, from <https://www.cdc.gov/diabetes/prevention/index.html>

Galaviz, K. I., et al., (2015). *Lifestyle and the Prevention of Type 2 Diabetes: A Status Report*. American Journal of Lifestyle Medicine, 12(1), 4–20. <https://doi.org/10.1177/1559827615619159>

Lindberg, S. (2020, June 17). *The cost of Type 2 Diabetes: Medication, Supplies, and more*. Healthline. Retrieved December 9, 2022, from <https://www.healthline.com/health/cost-of-diabetes>

National Addiction & HIV Data Archive Program (2003). *EDUCA: EDUCATION LEVEL*. The Regents of the University of Michigan. Retrieved December 9, 2022, from <https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/EDUCA?archive=NAHDAP>

National Addiction & HIV Data Archive Program (2003). *AGEG5YR: FOURTEEN-LEVEL AGE CATEGORY [Calculated Variable]*. The Regents of the University of Michigan. Retrieved December 9, 2022, from <https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/AGEG5YR?archive=NAHDAP>

NDPP. (n.d.). *National DPP Customer Service Center*. Centers for Disease Control and Prevention. Retrieved December 9, 2022, from <https://nationaldppcsc.cdc.gov/s/article/FAQ-Cost-of-the-Program-1526419435855#:~:text=On%20average%2C%20the%20cost%20per,who%20are%20uninsured%20or%20underinsured>.

Ou, H.-T., et al., (2016). *Life Expectancy and Lifetime Health Care Expenditures for Type 1 Diabetes: A Nationwide Longitudinal Cohort of Incident Cases Followed for 14 year*s. Value in Health, 19(8), 976–984. <https://doi.org/10.1016/j.jval.2016.05.017>

Poisal, J. A., et al. (2022, March 28), *National Health Expenditure Projections, 2021 - 30: Growth To Moderate As COVID-19 Impacts Wane*. Health Affairs, 41(4). <https://doi.org/10.1377/hlthaff.2022.00113>

Resource Center for Minority Data (2003). *INCOME2: INCOME LEVEL*. The Regents of the University of Michigan. Retrieved December 9, 2022, from <https://www.icpsr.umich.edu/web/RCMD/studies/34085/datasets/0001/variables/INCOME2?archive=RCMD>

Teboul, A. (2021). *Diabetes Health Indicators Dataset*. Kaggle.

<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

Xie, Z. (2019, September 19). *Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques*. Centers for Disease Control and Prevention. 16:190109. <http://dx.doi.org/10.5888/pcd16.190109>